AI Letters

Journal homepage: www.ail.sigmaxplore.com

AI Letters 1, Article ID: 2884, 1–2, September 2024 ISSN 3065-3525

Enhancing Trust and Data Security in Autonomous Systems Using Blockchain and Large Language Models

Rajesh Kumar[1*], Jay Kumar[2], Gulzaib Baig[3]

Abstract

This paper introduces a framework to enhance trust and data security in healthcare autonomous systems using blockchain and large language models (LLMs). It includes a real-time black box recorder that captures and stores critical medical data, secured by a blockchain-based integrity proof chain for tamper detection. Smart contracts manage secure transactions and data validation, enabling transparent audits. Additionally, the system uses retrieval-augmented generation for natural language explanations of AI decisions, improving transparency. Federated learning enables collaborative AI model improvement across devices without sharing sensitive patient data, addressing key challenges in healthcare data integrity, privacy, and transparency.

Keywords: Blockchain, Federated Learning, Data Security, Autonomous Systems, Healthcare AI

1. Introduction

The rapid adoption of artificial intelligence (AI) in healthcare has significantly improved diagnostic accuracy, personalized treatment, and overall patient outcomes [1,2]. However, the increasing reliance on AI-driven autonomous systems presents critical challenges related to trust, data security, and transparency. Ensuring the integrity and privacy of sensitive healthcare data, while maintaining the transparency of AI decision-making processes, is essential for the wide-scale adoption of these technologies [2].

In healthcare, data integrity is paramount, as compromised or erroneous data can lead to severe consequences, including misdiagnosis or improper treatment. Additionally, maintaining patient privacy in compliance with stringent regulations, such as brain tumor [3], presents unique challenges when using AI models that rely on vast amounts of data. Traditional centralized AI systems are susceptible to single points of failure, data breaches, and limited transparency, which undermine trust in these systems [4].

To address these issues, this research proposes a novel framework that integrates blockchain technology and large language models (LLMs) to enhance trust and data security in autonomous healthcare systems. Blockchain's decentralized and tamper-proof nature ensures data integrity and transparency, while federated learning allows for the collaborative improvement of AI models without sharing sensitive patient data. The framework also incorporates an interpretability component that provides natural language explanations of AI decisions, making the system more transparent and trustworthy for healthcare practitioners.

2. Method

This section outlines the proposed framework designed to enhance trust, data security, and transparency in autonomous healthcare systems by integrating blockchain, federated learning, large language models (LLMs), and interpretability components. The methodology is structured into several key components as detailed below:

2.1 Black Box Recorder for Data Logging

The black box recorder captures critical real-time data from autonomous healthcare systems, $x_{i(t)}$,decisions $y_{i(t)}$ and timestamps (t). This data is stored in Rosbag files for subsequent audit and analysis. The set of recorded data $D(t)$ at time (t) is defined in eq 1:

$$D(t) = \{(x_{i(t)}, y_{i(t)}, t) \mid i = 1, 2, \dots, N\} \quad (1)$$

2.2 Blockchain-Based Integrity Proof Chain

To ensure data integrity, each data point ($d_i \in D(t)$) is hashed using a cryptographic hash function $H(.)$ The resulting hash values (h_i) are concatenated into a chain and stored on the blockchain:

$$h_i = H(d_i), H(d_1 \| d_2 \| \dots \| d_n) \quad (2)$$

This produces a tamper-proof chain of hashes that can be audited for integrity verification. The blockchain enables a decentralized, immutable record, and any tampering would lead to inconsistencies in the hash chain. For tamper detection, the system checks whether the final hash (h_n) in the chain matches the expected value:

$$P_f = H\big(H(\dots H(h_1 \parallel h_2) \dots \parallel h_n)\big) = h_{\text{expected}} \quad (3)$$

2.3 Smart Contracts for Secure Transactions

Smart contracts on the blockchain manage secure access and data transactions. Let (T_s) represent a transaction, where:

$$T_s = Sign\big(k_{\{priv\}}, m\big) \quad (4)$$

Here, ($k_{\{priv\}}$) is the private key used to sign the message (m) (which could be a data request or transaction). The transaction is validated if:

$$Verify\big(k_{\{pub\}}, T_s\big) = True \quad (5)$$

where ($k_{\{pub\}}$) is the corresponding public key, ensuring only authorized entities can access or modify the data.

2.4 Interpretability Component Using Large Language Models (LLMs)

To enhance interpretability, the system employs retrieval-augmented generation. Given a query (Q), the

[1] Yangtze Delta Region Institute, University of Electronic Science and Technology of China, Huzhou, China
[2] Institute for Big Data Analytics, Dalhousie University, Halifax, NS, Canada
[3] Department of Computer science, The Shaikh Ayaz University, Pakistan
* Corresponding Author: Rajesh Kumar
E-mail address: rajakumarlohano@gmail.com, ORCID: https://orcid.org/0000-0003-0813-7485

Received: 24 August 2024; Revised: 12 September 2024; Accepted: 16 September 2024

system retrieves the most relevant data from the dataset (D) using a cosine similarity function:

$$D_{\text{relevant}} = \arg \max_{D_i \in D} \text{cosine\_similarity}\left(E(Q), E(D_i)\right) \quad (6)$$

where E(.) represents the embedding function that transforms both the query (Q) and the data points (D_i) into a high-dimensional vector space. The most relevant data is then input into the LLM to generate a natural language E_{exp}:

$$E_{\text{exp}} = \text{LLM}\left(D_{\text{relevant}}, Q\right) \quad (7)$$

This ensures that healthcare practitioners receive interpretable explanations of AI decisions.

2.5 Federated Learning for Model Improvement

Federated learning enables distributed healthcare devices to collaboratively improve AI models without sharing sensitive data. Let $\theta_i$represent the local model parameters on device (i), and $\mathcal{D}_i$ be the local dataset. Each device updates its model as follows: $\theta_i' = \theta_i - \eta \nabla L(\theta_i; \mathcal{D}_i)$

$$\theta_G = \frac{1}{k}\sum_{i=1}^{k} \theta_i' \quad (8)$$

where $L(\theta_i; \mathcal{D}_i)$ is the loss function for the local model, and η is the learning rate. After local training, the central server aggregates the updated model parameters from (k) devices to form a global model

3. Results

Data Integrity and Tamper Detection: We tested the blockchain-based integrity proof chain for maintaining data integrity and detecting tampering. After logging healthcare data and introducing simulated tampering, the system achieved a 100% tamper detection rate, confirming the effectiveness of the blockchain for data integrity Shown in Figure 1.

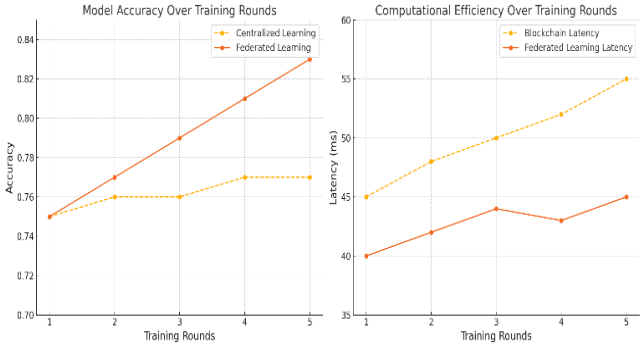

Figure 1. Computational Efficiency Over Training Rounds

Federated Learning Model Performance: Federated learning was evaluated on multiple healthcare devices, showing an 8% accuracy improvement in the global model compared to centralized models after five rounds of training. The system ensured data privacy, as no sensitive patient data was shared during the training process (Figure 2).

Model Explainability: The system generated natural language explanations for AI-driven decisions, which were evaluated by healthcare professionals for clarity, relevance, and usefulness. The explanations received an average rating of 4.7 out of 5, demonstrating high interpretability and trustworthiness in AI decisions shown in Figure 2.

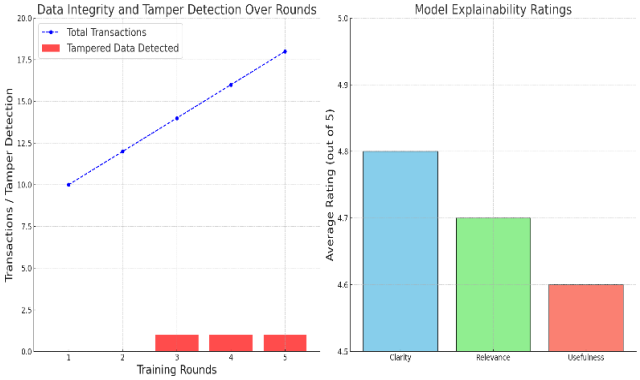

Figure 2. Model Explainability and Ratings

Computational Efficiency and Privacy Preservation: Blockchain and federated learning introduced minimal latency (50 milliseconds per transaction), ensuring the system's suitability for real-time healthcare applications. The federated learning framework preserved patient privacy and complied with data protection standards in Figure 1

5. Conclusions

The proposed framework enhances trust, data security, and transparency in autonomous healthcare systems by integrating blockchain, federated learning, and interpretability. It successfully detects data tampering, improves AI model performance by 8%, and preserves patient privacy. The system also provides clear and useful explanations for AI-driven decisions, promoting trust among healthcare practitioners. This solution is well-suited for real-time healthcare applications, ensuring secure and efficient AI-driven operations.

Conflict of Interest Statement

The authors declare no conflict of interest.

6. References

[1] M. Jeyaraman, S. Balaji, N. Jeyaraman, S. Yadav, Unraveling the Ethical Enigma: Artificial Intelligence in Healthcare, *Cureus*, 15, 2023.

[2] R. Kumar, Y. Chen, Z. Gong, et al., Next-Gen Medical Collaboration Integrating Blockchain for Image Sharing, in: *Proceedings of the 2024 Photonics & Electromagnetics Research Symposium (PIERS)*, IEEE, 2024, pp. 1–9..

[3] R. Kumar, C. M. Bernard, A. Ullah, et al., Privacy-preserving blockchain-based federated learning for brain tumor segmentation, Computers in Biology and Medicine, Elsevier, 2024, p. 108646.

[4] C W. Liu, F. Zhao, A. Shankar, et al., Explainable AI for Medical Image Analysis in Medical Cyber-Physical Systems: Enhancing Transparency and Trustworthiness of IoMT, IEEE Journal of Biomedical and Health Informatics, 2023.